2018-02-01

'Wrong' answers, right response: learning from randomised controlled trials when you don't get the results you were hoping for

Axford, Nick

BIG LOTTERY FUND

realising ambition

‘Wrong’ answers, right response:

learning from randomised controlled trials when you don’t get the results you were hoping for

Realising Ambition
Programme Insights: Issue 11

NATIONAL LOTTERY FUNDED
catch 22
DARTINGTON SERVICE DESIGN LAB
substance.
THE YOUNG FOUNDATION

Think Piece

Programme Insights: This series of Programme Insights shares reflections, learning and practical implications from Realising Ambition, a £25m National Lottery funded programme established by the Big Lottery Fund to support the replication of evidence-based and promising interventions designed to improve outcomes for children and young people and prevent them from entering the youth justice system.

Rather than writing a long evaluation report at the end of the five-year programme, this series has provided information about Realising Ambition in bite size chunks. This, the eleventh issue, summarises what we think are the key learning points to have emerged from undertaking Ramdomised Controlled Trials within the programme. It is a 'think piece' – qualitative reflections from the team. Words highlighted in blue are defined in the glossary.

About us: The Realising Ambition programme is managed by a consortium committed to improving outcomes for children. It is led by Catch22, alongside the Dartington Service Design Lab, Substance and The Young Foundation.

Realising Ambition Programme Insights: Issue II

Introduction

Realising Ambition was established to give young people positive opportunities to develop their potential and ambitions, and to build evidence and understanding about what works in preventing youth anti-social behaviour and crime. National Lottery funding from the Big Lottery Fund allowed the replication and support, where appropriate, for scaling of 25 proven and promising services across the UK. Funding was focused on improving outcomes regarded as indicative of, or precursors to, offending.

The Realising Ambition programme also provided significant input to strengthen the funded organisations and support for the design and delivery of funded services. A further feature of the programme was the systematic collection of data on implementation and impact.

Our fourth Programme Insight (available here) describes Realising Ambition's approach to routine outcomes monitoring to support service improvement efforts, and an example of aggregated data is illustrated in Table 1 below. Whilst a control group is absent, these data do indicate that the general movement in outcomes falls in line with expectations: appearing to improve for targeted early intervention services (more so for the more intensive and expensive services); and with universal prevention services successfully maintaining a stability in outcomes when they may otherwise be expected to deteriorate.

The reach of the programme was significant, with over 163,000 young people across the UK benefiting from services funded as part of Realising Ambition. Furthermore, 17 of the services that were funded continue to be delivered post-programme.

Three randomised controlled trials ('trials') were funded by the programme in order to help build the evidence base for promising social interventions developed in the UK for children and young people.

The decision to include trials was not taken lightly. They bring with them significant ethical and methodological challenges, notably the need to recruit sufficient participants and ensure the control group is treated fairly. They also present substantial practical and financial demands.

Motivations for stakeholders to engage in trials can range from a desire to contribute to the wider evidence base, through to wanting concrete 'proof' that a given service 'works'. Realising Ambition was underpinned by an appetite to learn about service effectiveness and use the findings to improve services.

"We wanted find out more about our service and the difference it made - and use this to help us drive development. We've had a longstanding commitment to evaluation so this seemed like the next step, and of course we wanted to add to the evaluation evidence base for mentoring more broadly."
Geethika Jayatilaka, Chance UK

Table 1: Movement in beneficiary outcomes

Category	Project*	Proportion Improved	Proportion Stable	Proportion Deteriorated
School-based prevention	LST	45%	26%	30%
	PACS	37%	37%	26%
Family Early Intervention	SFP 10-14	26%	49%	25%
	BANG	53%	22%	25%
Intensive Family Support	FFT	49%	22%	29%
	MST	56%	24%	20%

** LST = LifeSkills Training; PACS = Positive Assertive Coping Strategies; SFP 10-14 = Strengthening Families Programme 10-14; FFT = Functional Family Therapy; MST = Multi-Systemic Therapy.*

Having committed to conducting trials, Realising Ambition undertook a transparent and rigorous selection process. All delivery organisations were invited to submit an expression of interest. The selection was informed by:

- a decision to focus on services that had not been evaluated in a trial before, thereby providing an opportunity to add to the evidence base on promising interventions;
- a desire to achieve a balance between community-based and school-based services (reflecting the broad range of services within the Realising Ambition portfolio and to widen the contribution to the UK evidence base);
- a geographic spread (so far as possible); and
- the feasibility of undertaking a trial (Box 1).

Box 1: Key requirements for a service and a delivery organisation to be trial-ready:

- The service is tightly defined and delivered consistently (clear logic model supported by previous research, processes for ensuring consistent replication / delivery, well-defined target group and clear target group criteria applied consistently);
- There is senior-level and board support in the respective organisation, partly due to the inevitable demands placed on staff time of undertaking such a rigorous process (this included a willingness to develop capacity for staff to support the evaluation);
- Adequate numbers of participants can be engaged in the trial (capacity to serve a large enough number of children or parents, and sufficient demand for the service to permit randomisation on or off the service);
- Sufficient time within the parameters of available funding exists for the service to run in its entirety (covering recruitment and delivery) and for data to be collected;
- The possibility of randomisation (i.e. participants may be placed on a waiting list or receive services as usual and others may be assigned to receive the service being trialled).

Four services were selected for a trial on the basis that they best met the criteria. After further work, two ended up being main trials (Chance UK and Malachi), focusing on service effectiveness.

One (Ariel) became a pilot trial with an in-built feasibility study, because it was felt that the service would benefit from further development and because the numbers of schools and children required for a fully powered trial were greater than those engaged at the time. As such, the questions that needed answering here were less about effectiveness and more about whether the service was *likely* to have a positive effect, and if not how it could be modified and whether changes were needed to the way it was implemented. The pilot trial sought to assess the possibility and value of conducting an RCT. The fourth service selected for a trial (YMCA's Plus One Mentoring Service) did not ultimately proceed to a trial (see Programme Insight 6). [1]

All three organisations that participated in the trials received additional support focusing on their logic models, the component parts of their services, fidelity monitoring and strategies for recruiting participants. The services that were evaluated, together with study aims and methods, are summarised in Box 2.

Box 2: Summary of trials

Chance UK
Chance UK's mentoring programme is targeted at children aged 5-11 years with challenging behaviour and emotional problems at school and at home. The trial sought to recruit and randomly allocate 246 eligible children: intervention (n=123) and control (n=123). Data were collected at three points: baseline, midpoint (9 months) and endpoint (16 months). The primary outcome was the Total Difficulties score of the parent-rated Strengths and Difficulties Questionnaire (SDQ) at endpoint. Secondary outcomes included SDQ subscales (parent- and teacher-completed), the (parent-completed) Eyberg Child Behaviour Inventory (ECBI) and child-completed scales on self-perception and ability to pursue one's goals (for children aged 9 years and over). Data were also collected from Chance UK on aspects of implementation fidelity and from parents on other (non-Chance UK) services used by intervention and control participants.

Malachi
Malachi's "Inspiring Futures" is a therapeutic parenting group programme for parents of children aged 6-11 years with behavioural and emotional difficulties, with follow-up additional one-to-one sessions for selected parents. The study sought to recruit 248 children and randomise them to intervention (n=124) and control (n=124) conditions. Data were collected at three points: baseline, post-group sessions (16 weeks) and post one-to-one sessions (32 weeks). The primary outcome was the Total Difficulties score of the parent-rated Strengths and Difficulties Questionnaire (SDQ) at 32 weeks. Secondary outcomes included the SDQ subscales, the Eyberg Child Behaviour Inventory (ECBI), the Alabama Parent Questionnaire (focusing on parenting practices), the Ways of Coping Questionnaire and the Adult-Adolescent Parenting Inventory (AAPI) subscale on parent empathy. Data were also collected on implementation fidelity (from session checklists completed by facilitators and independent observation of video recordings of selected sessions) and, from parents, on other (non-Malachi) services used by intervention and control participants.

Ariel
Ariel's "Face Up" is a universal school-based programme for children aged 11-16, designed to promote healthy dating relationships and prevent psychologically abusive behaviour. An initial feasibility study led to adjustments to the intervention, including new curriculum content and the addition of a whole-school element. The pilot trial of this revised version, with a built-in feasibility study, involved five schools: intervention (n=3) and control (n=2). It focused on: (i) recruitment strategies; (ii) fidelity and acceptability of intervention; and (iii) suitability of data collection methods. Methods included: (i) observation of training, selected lessons and the whole-school element; (ii) lesson records (teacher self-completion); (iii) interviews with teachers; and (iv) focus groups with pupils; and (v) an online pupil survey comprising measures relating to psychological abuse in dating relationships and covering knowledge, attitudes, skills, experience and likely action in the event of experiencing or observing abusive behaviour.

Results

In order to ensure that the full results can be published in peer-reviewed scientific journals, it is only possible at this point to precis the results from the three studies.[2]

Data analysis is ongoing, but we can say now that there were no effects on the primary outcome in either main trial. In the Chance UK trial, most outcomes improved over time in both intervention and control groups, and although this mostly favoured the intervention group the differences were small and none were statistically significant. There was a small statistically significant positive effect on one secondary outcome in the Malachi trial at the mid-point but this had disappeared by the end of the study (a few months later). There was no evidence in either trial that selected sub-groups of children did better than others on the primary outcome (e.g. boys vs. girls). There was also no effect on the primary outcome in the Chance UK trial when the length of intervention received was taken into account. [3]

The Ariel studies, meanwhile, provided much learning about Face Up, particularly in relation to the process of school recruitment, teacher training and implementation. This has led to significant changes to the programme and informed how Ariel has developed subsequent interventions.

The remainder of this Programme Insight concentrates primarily on the two main trials, although the pilot trial is included to show how results from all three evaluations can be used to improve the services concerned. Throughout we draw on our own reflections as well as those of the delivery organisations and other stakeholders. [4]

Responding to null effects

We arguably learn as much, if not more, from a null result as we do from a positive one. Consequently we have sought to avoid unhelpful reactions to results we did not hope see. Often, an overall null effect in a trial leads to authors cherry-picking any favourable results and giving them undue prominence, or service implementers querying the veracity of the methods used, such as study size or choice of measures, suggesting these are responsible for the failure to detect a positive impact. There is also a tendency for the academic field to be less interested in null effects – manifested, often, in a failure to publish results.[5]

Realising Ambition seeks to model a more positive and thoughtful response to finding null effects. Far from a null effect from one high-quality trial necessarily equating to a failed service, we believe that it can point to valuable learning; indeed, if the results are used and acted upon, it can be part of a normal and healthy process of service improvement. [6] Reflecting the Realising Ambition ethos of "improving as well as proving", therefore, we are open here about the results and have reflected collectively on what can be learned – both to improve the services evaluated but also for the wider field.

That isn't to say that the results are not disappointing. Moreover, having undertaken all due process, the outcomes were not anticipated. We were confident that services and delivery systems were robust; indeed, they were scrutinised as part of the selection process. And outcome data collected previously by the respective organisations, together with personal observation, suggested that the services made a positive difference to children and young people.

Neither trial is without limitations, which we will acknowledge in the full reports. Programme Insight 6 (available here) reflected on the process of conducting the trials and the main challenges we encountered – for instance around recruitment and monitoring fidelity. Broadly, however, we are confident that both trials were well designed and executed, and as such have accepted the results.

Learning about the services

In the Chance UK trial, the complexity of children's needs, which were well into the clinical range on average[7], may help to explain results. Accordingly, the service is being adapted to address more fully issues in children's lives that might be contributing to their behavioural and emotional difficulties. For instance, work has already started on strengthening the parenting element of the

service. Malachi, meanwhile, is seeking to boost implementation fidelity by ensuring that more parents attend more group sessions and increasing the extent to which group facilitators deliver the required content as intended.

It is unclear whether these are the only factors that account for the results in the outcomes measured in the trials, so in both cases there is a case for revisiting service design in the light of up-to-date evidence in the field of youth crime and anti-social behaviour. This would include reviewing the respective logic models in greater depth than was deemed necessary or possible prior to the Realising Ambition trials.

It makes sense to introduce adaptations carefully and test them quickly as part of an iterative process. This method, called rapid cycle testing, focuses initially on implementation (can the changes be delivered?) and acceptability (do the people delivering the service and those receiving it engage with and like the changes?). Testing of outcomes can follow, starting with change over time followed by more robust tests subject to early results. Whilst still early in concept, the Dartington Service Design Lab and others are developing thinking and practice in this regard.

"The long timeframe associated with the trial was inflexible; it didn't respond to the changes that were happening in the political environment, in the educational environment and in our schedules. By the time you've collected evidence, the agenda's moved on. In relation to those problems that are changing and evolving quickly, you've got to use a much more rapid cycle approach to building those evidence bases". Paul Ainsworth, Ariel Trust

Learning for the field

What can be learned from the Realising Ambition trials for the field of early intervention in children's social care and youth offending, particularly in respect of the place and conduct of trials?

"Improving via proving"

The process of preparing for and conducting a trial can help to strengthen the service being evaluated.[8] In the Realising Ambition trials we strengthened the respective logic models (albeit less for Chance UK and Malachi than for Ariel), modified eligibility criteria (to better target families with sufficient levels of need), co-developed fidelity monitoring tools and assisted with recruitment.[9]

The exacting nature of trials arguably means that these aspects of design and implementation are improved more in the context of an RCT than they might be otherwise. That said, the strictures imposed by trials can also make service delivery more challenging. The numbers required in the Malachi trial, for instance, led to modifications to the normal recruitment method, rapid turnaround from recruitment to parent groups starting and a larger-than-usual number of facilitators, all of which may have affected fidelity. Similarly, Chance UK reported that the demands of the trial meant that some of its normal development processes were put on hold.

Of course, improving services through the trial process is not the primary reason for doing trials. Rather, we do them to find out if a service or approach is effective in improving specified outcomes. This means that we should only do them if we are genuinely interested in the results and in acting on them. Through both process and results, then, trials facilitate what we might call "improving via proving". Accordingly, there is a case for embedding trials in a developmental process rather than seeing them as the end in themselves, which leads to the next point.

"The journey has led us to a whole load of different places mindset-wise, design of new intervention-wise, dreaming of new ideas, innovation around things. Emphasis on training has come into this organisation like never before because of the fidelity findings. Holding on to the improvements that we've made organisationally has been a huge positive". Laura Evans, Malachi

Incentivising trials

Although it should be the norm for organisations to test whether beneficiaries are getting what it says 'on the tin', it isn't. It takes courage for an organisation to allow its service to go through an RCT, especially if it is core business. To that end Chance UK, Malachi and Ariel should be applauded.

Just as receiving a service that has not been tested rigorously exposes children and families to risk, so being involved in a trial raises the risk of the respective delivery organisation *appearing* – in the case of null or negative effects – to be in a worse position than before.[10]

Understandably, this possibility can be a disincentive to participating in a trial. We need to guard against this. For example, a guaranteed "scale-up fund" (if the service is found to be effective) or "improvement fund" (if the service is not found to be effective) would give innovators some degree of security. More radically, if funding were *generally* more forthcoming to services with demonstrable evidence of effectiveness (unless there was good reason for this not to happen) then experimental evaluations could become the norm, increasing incentives to undertake them.

"Coming in at the end of all this, I think the thing that stands out for me is that there was a lot of work in the run-up to the RCT, a lot of preparation. I think for me there's been a very hard edge which is: we've done the RCT, we've had the result, and there's the full stop at the end of the sentence. Actually if we do see this as a developmental process, and I think we should, not just us as an organisation but also funders and the sector as a whole, then there actually should be some support and planning the other side of this as well". **Geethika Jayatilaka, Chance UK**

Aligning evaluation methods and purpose

Evaluation methods need to be aligned with the stage of service development. This works in two ways: it is important not to do trials too early and set services up to 'fail'; but equally it is unhelpful for established services to avoid them while using pre-post studies, KPIs and qualitative feedback as a basis to advertise effectiveness. Broadly we think we got this right in Realising Ambition. Both the Chance UK and Malachi services had previous evaluation studies, were assessed for their readiness for a trial and received considerable support with service development and implementation. They were also well established in their local areas and had been serving significant numbers of children and families annually for several years. Ariel, by contrast, was considered to need further development, hence the pilot trial.

Even so, we have questioned whether Realising Ambition could have done more to maximise the chances of finding positive effects. For instance, could we have assessed more carefully whether the Chance UK service aligned with the evidence on factors associated with effective mentoring?[11] And could we have done more to ensure that Malachi's service reflected best evidence and recruited and retained enough parents? Concerted efforts were made pre-trial to address these issues, and to troubleshoot where problems arose, so the answers are not immediately obvious. Moreover, it is not necessarily possible or desirable to modify established services during trials.

What about the investment? Are RCTs worth the money? They tell us something about the service being trialled and potentially contribute to wider knowledge (because results from trials in a given field get added up to give us a sense of what works overall). [12] Some commentators would argue that learning from trials is limited because they only tell us what works in a given context at a given time, making it hard to generalise.[13] While we have some sympathy with that perspective, in evidence-based commissioning and practice we should reflect on those things anyway (i.e. finding evidence in a trial of effect / no effect isn't deterministic), and where the evidence from several contexts tells a consistent story it should boost our confidence that a particular service or category of services is transportable. [14]

Making trials more revealing

A lot of trials are 'thin', by which we mean that they tell us about a service's impact on the target outcome(s) and little else. This can leave service developers scrabbling for clues about what to do next, especially if the results are equivocal or disappointing. Generally, we need to get better at gathering data and conducting analyses that help make sense of the results regarding service impact: we need 'thick' trials.

Two areas for improvement involve recording fidelity in more depth than is common, and collecting more robust data on what control group participants receive. We think we did both of these reasonably well in the Realising Ambition trials.

Fidelity was recorded through practitioner self-report in both main trials and strengthened in the case of Malachi by independent observation (via the coding of video-recorded group sessions). However, adherence to core content could have been measured better in both cases by capturing in finer detail what was delivered relative to the design. We also sought to explore the relationship between fidelity and outcomes in both trials, although methodologically this is challenging and to date arguably we have been able to do less than the respective organisations would have liked. [15]

In both main trials we also gathered considerable data on other services used by children and parents in both intervention and control groups. That said, ascertaining the degree to which other services received by the participants in the control group resemble the service being trialled requires more conceptual and methodological work in the field generally.

With additional resource we could also have usefully conducted a more thorough process evaluation (as we did for Ariel). Exploring issues in delivery, aspects of context and the influence of the wider system would enhance our understanding of the outcomes findings. The analyses of this kind of data should be impact dependent. For example, if a trial finds a positive effect we are likely to be interested in identifying barriers to, and facilitators of, implementation and scale. Whereas a null effect should lead to an assessment of why this is and how the intervention can possibly be improved.

A further future addition to trial design could be 'mediation analysis'. This would help us understand why or how interventions work by testing whether the hypothesised change mechanisms (as set out in the logic model) materialise in practice. If a service is effective in terms of specified outcomes it is useful to know why it worked, in part because others can borrow that mechanism. There is a tendency to leave this 'black box' untouched, which is a shame because we learn much less than we might if we take time to open it and peer in. The Realising Ambition trials took steps in this direction by testing the effects on *intermediate* outcomes (which for the most part also showed no effect).

"We want to fully unpick those findings so that we've got a much deeper understanding of them in all of their intricacies. I still believe that the programme makes a difference, I've seen it with my own eyes. Whether or not we've measured the difference it does make, well, the findings would debate otherwise, but I don't want to be in a position of throwing the baby out with the bathwater. So what I need to do is work out: what's the bathwater here and what's the baby?"
Laura Evans, Malachi

Valuing control groups

Where used ethically, control groups are extremely valuable because they can help us to see whether or not work with children and families is beneficial. In the case of Chance UK, participants in the intervention group improved on average on most outcomes over time: their behaviour got better and their emotional difficulties reduced. Mentors would likely have observed changes in the right direction and attributed those improvements to their work with the children concerned. However, the same trends occurred in the control group. Although children in the intervention group improved slightly more on average than those in the control group on most outcome measures, any differences between groups were very small, and none were statistically significant.

We emphasise the value of a control group because there is growing interest in the idea of 'improving' rather than 'proving' [16], where the former is more likely to look at change in outcomes over time – if at all – and the latter is shorthand for studies that have intervention and control groups (including trials). It should not be a case of either/or. Control group studies can – and some would say must – be integral to service improvement.[17]

Getting ready for results

Arguably the field needs to get better at preparing stakeholders for the results from trials, and being ready to learn from those results. In Realising Ambition, a year before results were due, we held a half-day workshop with all three organisations and considered the following:

- how well the research was proceeding and if there were any concerns about it (for example about measures);
- how well service delivery was proceeding (for instance, any observed issues with implementation fidelity or other delivery processes);
- the possible "dark logic" of the respective services (in other words, how might they have adverse effects on target or other outcomes, and was there any sense of this happening);[18]
- the extent to which various contextual factors (social, political, cultural and organisational) might be influencing the effectiveness of the interventions; and
- scenario planning i.e. how would the respective organisations respond to different patterns of findings regarding service effectiveness (in the crudest sense: effective, not effective, harmful, mixed).

This was a useful exercise, and one that to our knowledge is not common in trials, although with hindsight we could have scrutinised available fidelity data more fully at this point.[19] It is also worth reflecting that whereas extensive service refinement support was available to the organisations before and to some extent during the Realising Ambition trials, we had not planned any such activity after the trials ended.

Rethinking service design and evaluation

The results of the Realising Ambition trials need to be seen in the context of evidence from other trials of prevention and early intervention programmes for children and families. Although a simplification of a complex area, effects generally range from non-existent to small, and when interventions tested and found effective in one setting are replicated elsewhere the effects often disappear.[20] We rightly celebrate individual successes,[21] and small effects spread across large populations can add up, but when studies are examined closely for methodological quality and robustness of results, the overall picture can seem rather unsatisfactory.[22] In this respect, the Realising Ambition trial results are by no means aberrations.[23]

Playing into this scenario is a growing understanding of the complexity of the problems that we are often seeking to prevent or address. Viewed in this way, single interventions are simply events or 'disturbances' in a more complex system.[24] Some might ask what chance 12 months of volunteer mentoring or 12 parent group meetings stand in terms of achieving their specified outcomes in this context, and even question the value of conducting a trial on a single intervention.

In response, we would argue that individual interventions do not sit outside of the system: rather, they are part of it and can shape other parts.[25] An individual intervention or service to help people stop smoking, for example, helps to create a climate in which smoking is seen as undesirable and can pave the way for or reinforce public health interventions like smoking bans in pubs. There is also substantial evidence that discrete interventions *can* have positive effects – at least in the short-term.[26] Even so, with the goal of increasing service effectiveness there is a case for re-designing individual interventions to better address relevant systems. For example, there is evidence that school-based health interventions work better if they contain a family or community element also.[27] We might also focus more on aligning multiple parts of the system, including the discrete intervention, so that they reinforce one another.[28] Some implications of these ideas for service design and evaluation are explored in Box 3.

Box 3: New developments in intervention design and evaluation

In order to inform 'system-savvy' services the process of service design requires change. This might entail:

- using modelling techniques to explore how changes to one part of the system will affect others;
- constructing what we might call 'nuclear' and 'extended' logic models, where the latter extends the former's traditional focus on proximal risk and protective factors and target outcomes to encompass distal factors and potential side-effects (positive and negative);
- co-producing services in context with a range of stakeholders (especially service users); and
- building services to fit the simple rules of service systems. [29]

Changing how services are designed has implications for evaluation. Specifically, we need to:

- examine a wider range of outcomes, being alert to potential positive side-effects and unintended adverse outcomes;
- pay more attention to mediators (to explain why something does or doesn't 'work') and moderators (to explore for whom it 'works');
- gather more qualitative data from multiple stakeholders on the effects of the service, not just on the outcomes of interest but also on the system; and
- use methods that allow us to iterate as we discover (e.g. rapid cycle testing and adaptation) and triangulate different types of information from multiple sources.

What next for the three organisations?

After trial results were shared with the participating organisations a supportive process to explore implications for further service development and testing was initiated by the Realising Ambition consortium. Chance UK is focused on where the service can be 'dialled up' to maximise its effect. This includes developing further the parent component of the service. Malachi is concentrating on strengthening guidance and training for facilitators to boost implementation fidelity. Ariel's programme has been revised in the light of results from the feasibility study. Although Ariel's efforts are now more targeted at primary schools and include programmes aimed at preventing radicalisation, it is anticipated that the pilot study results will continue to shape programme design and implementation, including fidelity and outcome data monitoring.

Summary and conclusions

What may voluntary and community sector organisations that develop, deliver and adapt services to help prevent adverse outcomes, and those who fund and commission them, learn from Realising Ambition's experience?

The first and most obvious point is that consideration should be given in the first place to whether or not to do a trial. Assuming that the question focuses on impact, and quasi-experimental studies have been discounted, this means only doing a trial when the service is ready and there is genuine interest in whether it is effective. The organisations undertaking the trial and their funders must understand the implications, the financial investment and time required.

The second point is that trials need to be designed to optimise learning opportunities, regardless of the results. This means building in the resources and staff skills to collect good process and implementation data to inform scale-up or major revisions. Understanding the services received by the control group is crucial, as is exploring whether some sub-groups benefit disproportionately. Consideration should also be given to undertaking mediation analysis to explore the mechanisms through which the intervention is hypothesised to work.

Third, we need to foster a learning culture that trials can sit within, so that finding a null effect is accepted as part of a normal and healthy process of service development. This means being honest about results, taking steps to act on learning and protecting – or even rewarding – organisations that have the courage to allow their intervention to be subjected to a trial.

The fourth and arguably most radical conclusion is that we might need to rethink the design of services and how best to evaluate them. If we see services as events in systems it is likely to affect how we design them, the questions we ask about them and the methods of evaluation. Specifically, we will move from designing individual programmes to designing system changes and shift from asking whether a single service 'fixes' the problem to how it contributes to reshaping the system. In this case evaluating impact on a narrow set of outcomes is unlikely to be sufficient.

So, did we get these things right in Realising Ambition? Yes and no. On the decision to proceed to trial, we thought hard about which services to evaluate – the fact that Ariel didn't proceed to a main trial is an indicator of that – and sought to be candid with all stakeholders about the associated demands. We also invested significant resource in supporting those organisations that undertook a trial. On the other hand, although service development work was possible pre-trial, critics may argue we could have done more.

What of the trial designs? We collected reasonably good data on fidelity and what services the control groups received, which helped when considering reasons for the results. That said, the data on process were relatively limited in the two main trials. Certainly the much more extensive process data generated in the Ariel feasibility study and pilot trial permitted greater learning about programme development (though, importantly, not about effectiveness).

Moreover, there is no escaping the fact that these were individual services evaluated with a focus on a narrow set of outcomes. Until now this has been pretty mainstream but as we reflect on the success or otherwise of this general approach, and new insights from related fields (e.g. complex systems), we are having to re-think our approaches to service design and evaluation.

One final, ironic, thought: had the trial results been positive we would likely have reflected – and arguably learnt – far less.

Key Learning Points

- Consideration should be given in the first place as to whether or not a randomised controlled trial is the right approach to evaluating a service.
- Trials should only be undertaken if the questions focus on impact, quasi-experimental studies have been discounted and when the service is ready. The organisation whose service is being evaluated, and its funders, should understand the implications, time and financial investment required.
- Trials need to be embedded in a developmental process because through both process and results they facilitate what we might call "improving via proving".
- Trials need to be incentivised, for example through providing scale-up funding if the service is found to be effective or improvement funding if null effects are found.
- The field needs to get better at gathering data (e.g. on fidelity) and conducting analyses (e.g. on moderators) that help to make sense of results regarding service impact. This means building in the resources to do so, and developing staff skills to collect good data.
- Control groups should be valued more because they can help show if work with children and families is beneficial.
- Smarter preparation for trial results, and a commitment to learn from them, should entail considering in advance how to respond to different result scenarios.
- Null results are relatively common in trials of services for children and families. We need to reflect on the implications of this for service design and evaluation.

Glossary of Terms

- **Attrition**

The loss of participants from the study, typically defined as the number or proportion of participants who drop out.

- **Adherence**

the extent to which core intervention components are delivered (see also Fidelity).

- **Complex system**

a system comprising many components that may interact with one another.

- **Control group**

participants in an experimental or quasi-experimental study who do not receive the intervention that is being evaluated but are otherwise broadly similar to participants who do receive the intervention (i.e. the counterfactual).

- **Dark logic**

the mechanisms by which an intervention hypothetically has adverse effects on the outcome(s) of interest and potentially other outcomes.

- **Distal**

factors that are far or *distant* from the outcome of interest but nevertheless plausibly or empirically affect that outcome (see also Proximal).

- **Feasibility study**

examines the practicality of an intervention with a view to refining it. It looks at the acceptability of and engagement with the intervention as well as adherence in delivery and viability of implementation.

- **Fidelity**

whether an intervention is delivered as intended/designed, covering adherence (delivery of core components), exposure (delivery of the specified dose), quality (e.g. provider's preparation, attitude and engagement of participants), responsiveness (engagement of participants in the activities) and reach (in terms of the target group).

- **Improvement science**

finding out how to make changes to interventions and systems so that they perform better.

- **Mediation analysis**

statistical analysis conducted to identify mediators.

- **Mediators**

variables via which the intervention is hypothesised to or actually does impact on outcomes; for example, improved parenting skills might be the mediator of the impact of a parenting programme on children's behaviour.

Glossary of Terms

- **Meta-analysis**
a statistical method for combining the results from two or more studies.

- **Moderators**
variables according to which intervention effectiveness is hypothesised to vary or does vary – for example, gender would be a moderator if, say, girls did better on average than boys.

- **Null effects**
no effect on the outcome(s) of interest, on the basis that participants in the intervention group did no better or worse on average than participants in the control group.

- **Participants**
in the context of research, participants are individuals who agree (provide voluntary consent) to take part in a study, and should be distinguished from service users – in a trial, some but not necessarily all users of an intervention will be participants, and consenting individuals who do not receive the intervention because they are in the control group are also participants.

- **Pilot trial**
a randomised controlled trial in miniature, designed to test whether a full trial is feasible and worthwhile.

- **Primary outcome**
the outcome of primary interest; this will be the outcome that is used when determining the appropriate size for the trial, and to a large extent results on this outcome will determine whether or not the intervention concerned is deemed to be effective.

- **Promising**
denotes interventions that have not yet been evaluated in an experimental study but which are assessed as having the potential to improve target outcomes.

- **Proven**
refers to interventions with evidence of a positive impact from one or more experimental studies.

- **Proximal**
factors that are near to or in the *proximity* of the outcome of interest and plausibly or empirically affect that outcome (see also Distal).

- **Randomised controlled trial**
an experimental study in which participants are allocated to the study conditions or groups (e.g. intervention and control) at random (i.e. by chance alone).

- **Rapid cycle testing**
iterative testing of changes to an intervention with a view to improving the intervention.

Glossary of Terms

- **Recruitment**
the process of getting initial involvement and sign-up from participants in the study.

- **Replication**
refers to implementing in a new setting an intervention developed and tested elsewhere and/or finding in a new trial an effect that is comparable to the one found in the original trial.

- **Secondary outcomes**
outcomes in a trial that are of secondary importance (see also Primary outcome).

- **Services as usual**
refers to what services participants in the trial would typically receive in the absence of the trial; this often equates to the control condition (also referred to as 'treatment as usual')

- **System savvy**
interventions that comply with the simple rules of systems, meaning that they fit well and have the potential to become easily embedded in routine practice.

- **Systematic review**
a review of literature on a specific subject (e.g. effectiveness of a class of interventions) that uses explicit and transparent methods, follows a standard set of stages, and is accountable, replicable and updateable.

- **Thick trial**
a trial that evaluates the impact of the intervention on the outcome(s) of interest but gathers and analyses additional data to help explain the results.

- **Thin trial**
a trial that focuses exclusively on the impact of the intervention on the outcome(s) of interest, with little or no additional data or analysis to help explain the results.

- **Transportable**
a quality of an intervention – or a class of interventions – that (i) can relatively easily be delivered in a setting that is different to the one in which it was developed, and/or (ii) produces effects when tested in a new setting that are similar to those identified when tested in its original setting.

- **Waiting list**
refers to a trial in which the participants in the control group receive the intervention after intervention group participants have finished the intervention and all participants (intervention and control) have provided data post-intervention (or at the equivalent time point in the case of the control group participants).

Further Reading

Bonell, C., Fletcher, A., Morton, M., Lorenc, T. & Moore, L. (2012) Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Social Science & Medicine* 75 (12), 2299-2306.

Giangregorio, L. M. & Thabane, L. (2015) Pilot studies and feasibility studies for complex interventions. In Richards, D. A. & Hallberg, I. R. (Eds) *Complex Interventions in Health: An Overview of Research Methods*, London, Routledge.

Haynes, L., Service, O., Goldacre, B. & Torgerson, D. (2013) *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials*. London: Cabinet Office.

About the Authors

This Programme Insight (2018) has been written by Nick Axford, Shaun Whelan and Tim Hobbs. Nick is PenCLAHRC Associate Professor in Health Services Research at Plymouth University and a Senior Associate at the Dartington Service Design Lab. Shaun Whelan is Programme Manager of Realising Ambition, based at Catch22. Tim Hobbs is Director of the Dartington Service Design Lab. Nick Axford's time is supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South West Peninsula. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Acknowledgements

We gratefully acknowledge the helpful comments on earlier drafts from the following colleagues: Paul Ainsworth, Sarah Blower, Laura Evans, Frances Flaxington, Kate Green, Christine Hatt, Caroline Hopkins, Geethika Jayatilaka, Julian Lee, Eileen McKibbin, Louise Morpeth, Esther Rush, Janice Riddell and Matthew Riddell. Ben Hartridge (Dartington Service Design Lab) conducted the interviews with Ariel, Chance UK and Malachi.

Endnotes

[1]The CEO of YMCA Scotland and other senior staff were very supportive and enthusiastic about a trial. However, the PlusOne intervention was based on a social franchise model, so the RCT needed approval from franchise partners, many of whom felt the random allocation of young people to control and intervention conditions was deeply unethical. This, combined with issues about the number of participants needed to make the trial meaningful, contributed to the decision not to proceed with a trial. Instead, efforts were concentrated on further refining the intervention logic model and completing a breakeven analysis.
[2]Both trials were registered: further details about them can be seen at http://www.isrctn.com/ISRCTN47154925 (Chance UK) and http://www.isrctn.com/ISRCTN32083735 (Malachi).
[3]A similar analysis was not possible in the Malachi trial because the statistical criteria for undertaking it were not fulfilled.
[4] Some of the material comes from telephone interviews conducted with Paul Ainsworth (Director, Ariel Trust), Laura Evans and Julian Lee (co-CEO's, Malachi Trust), and Geethika Jayatilaka, Christine Hatt and Caroline Hopkins (respectively CEO, Quality Evaluation Manager and Senior Programme Manager, Chance UK). Reflections have further been informed by discussions at a seminar involving a wider group of Realising Ambition stakeholders held at the Young Foundation on 30th November 2017.
[5]Either because the authors do not submit the paper, or because journals have traditionally been less interested in publishing null effect trials (the two are related of course).
[6]For example, rthe Early Intervention Foundation writes this about the 'Not effective' (NE) rating in the evidence standards that underpin its Guidebook: "This rating should not be interpreted to mean that the programme will never work, but it does suggest that the programme will need to adapt and improve its model, learning from the evaluation. The best evidenced programmes have normally had null findings along the way to demonstrating proof of concept. Some developers with such evidence have terminated their programme, others are working out how to adapt and improve their model to respond to the evidence." http://www.eif.org.uk/eif-evidence-standards/
[7]According to the Total Difficulties score on the parent-rated Strengths and Difficulties Questionnaire: www.sdqinfo.com
[8]Axford, N., Berry, V. & Little, M. (2006) 'Enhancing service evaluability: lessons from a programme for disaffected young people', *Children & Society* 20 (4), 287-298.
[9]The methods for doing this are described in more detail in Programme Insight 6.
[10]We say 'appearing' because we think the organisations concerned are actually better off knowing the results from a robust evaluation of their respective interventions and being able to use this information to improve the interventions accordingly.
[11]See DuBois, D. L., Portillo, N., Rhodes, J. E., Silverthorn, N. and Valentine, J. C. (2011). How effective are mentoring programs for youth? A systematic assessment of the evidence. *Psychological Science in the Public Interest* 12 (2), 57-91.
[12]Systematic reviews of the effectiveness of a particular class of interventions include either a narrative synthesis or a statistical

synthesis (a meta-analysis) of the results of the evaluations they include – see Gough, D., Oliver, S. and Thomas, J. (2012) *An Introduction to Systematic Reviews*. London: Sage.

[13]Cartwright, N. and Hardie, J. (2012) *Evidence-based Policy: A Practical Guide to Doing It Better*. Oxford: Oxford University Press.

[14]For example, see Gardner, F., Montgomery, P. and Knerr, W. (2016) Transporting evidence-based parenting programs for child problem behaviour (age 3-10) between countries: systematic review and meta-analysis. *Journal of Clinical Child & Adolescent Psychology* 45 (6), 749-762.

[15]The type of participants who receive the full (or fuller) amount of an intervention tend to be different to those who do not (in terms of factors that are predictive of the outcomes). Common approaches to exploring the relationship between fidelity and outcome, whereby 'completers' and 'non-completers' within the intervention group are compared with each other, or completers in the intervention condition are compared with the entire control group, are misleading because they fail to account for the aforementioned differences between these groups.

[16]The type of participants who receive the full (or fuller) amount of an intervention tend to be different to those who do not (in terms of factors that are predictive of the outcomes). Common approaches to exploring the relationship between fidelity and outcome, whereby 'completers' and 'non-completers' within the intervention group are compared with each other, or completers in the intervention condition are compared with the entire control group, are misleading because they fail to account for the aforementioned differences between these groups.

[17]There is growing interest in improvement science – see, for example, Langley, G. J., Moen, R. D., Nolan, K. M., Nolan, T. W., Norman, C. L. and Provost, L. P. (2009) *The Improvement Guide: A Practical Approach to Enhancing Organizational Performance (Second Edition)*. San Francisco: Jossey Bass.

[18]Many efforts to optimise performance in a range of sectors are based on conducting and responding to the results of trials. For example, see Syed, M. (2015) *Black Box Thinking: The Surprising Truth About Success*. London: John Murray.

[19]This draws on the ideas set out in Bonell, C., Jamal, F., Melendez-Torres, G. J. and Cummins, S. (2015) 'Dark logic': theorising the harmful consequences of public health interventions. *Epidemiol Community Health* 69: 95-98.

[20]For example, although in the Malachi trial facilitator self-completion fidelity were monitored throughout, presenting an overwhelmingly positive picture, the task of viewing and coding the videos of selected sessions, which raised concerns about implementation, did not happen until towards the end of the trial.

[21]Sometimes referred to as the 'replication crisis'.

[22]For example, the Incredible Years parenting programme has been demonstrated to be effective in multiple diverse settings.

[23]For instance, the highly-regarded Blueprints for Healthy Youth Development project has reviewed over 1500 programmes and approved about 70 (c.5%): www.blueprintsprograms.com.

[24]This fact alone is arguably worth sharing by way of preparing organisations for trial findings.

[25]Rutter, H., Savona, N., Glonti, K. et al. (2017) The need for a complex systems model of evidence for public health. *The Lancet*, 390, Issue 10112, 2602-2604.

[26]Sniehotta, F. F., Araújo-Soares, V., Brown, J. et al. (2017) Complex systems and individual-level approaches to population health: a false dichotomy? *The Lancet Public Health* 2 (9), e396-e397.

[27]As evidenced by the interventions listed on www.blueprintsprograms.com or rated at Levels 3 or 4 on www.guidebook.eif.org.uk.

[28]Langford, R., Bonell, C. P., Jones, H. E., Pouliou, T., Murphy, S. M., Waters, E., Komro, K. A., Gibbs, L. F., Magnus, D., and Campbell, R. (2014), The WHO Health Promoting School framework for improving the health and well-being of students and their academic achievement, *Cochrane Database of Systematic Reviews* 2014, Issue 4.

[29]Hanleybrown, F., Kania, J. and Kramer, M. (2012) *Channelling Change: Making Collective Impact Work*. Stanford Social Innovation Review. http://ssir.org/articles/entry/channeling_change_making_collective_impact_work [Accessed 18th January 2018]

[30]http://preventioncentre.org.au/blog/enacting-change-in-complex-systems-why-theory-matters/ [Accessed 18th January 2018]

Find out more

catch 22

realisingambition@catch-22.org.uk

catch-22.org.uk/realising-ambition

DARTINGTON SERVICE DESIGN LAB

tim.hobbs@dartington.org.uk

dartington.org.uk

substance.

neil.watson@substance.net

substance.net/case-studies/realisingambition

THE YOUNG FOUNDATION

james.teasdale@youngfoundation.org

youngfoundation.org


BIG LOTTERY FUND
NATIONAL LOTTERY FUNDED

catch 22

DARTINGTON
SERVICE
DESIGN LAB

substance.

THE
YOUNG
FOUNDATION